Tạp chí Tin học và Điều khiển học, T.29, S.1 (2013), 211–218

MỘT LƯỢC ĐỒ THỦY VÂN CƠ SỞ DỮ LIỆU QUAN HỆ VỚI DỮ LIỆU PHÂN LOẠI

LƯU THỊ BÍCH HƯƠNG[1], BÙI THẾ HỒNG[2]

[1] *Khoa Công nghệ thông tin, Trường Đại học sư phạm Hà Nội 2*

[2] *Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học & Công nghệ Việt Nam*

Tóm tắt. Thủy vân là một trong các giải pháp hữu hiệu dùng để bảo vệ bản quyền và sự toàn vẹn cho các cơ sở dữ liệu quan hệ trong môi trường internet. Trong bài báo này chúng tôi đề xuất một lược đồ thủy vân mới có thể phát hiện và khoanh vùng các giả mạo trong các cơ sở dữ liệu quan hệ với dữ liệu phân loại. Kỹ thuật thủy vân này không làm thay đổi giá trị các bộ trong cơ sở dữ liệu mà chỉ thay đổi một cách bí mật và an toàn thứ tự của các bộ trong cơ sở dữ liệu.

Từ khóa. Thủy vân, cơ sở dữ liệu, dữ liệu phân loại.

Abstract. Watermarking is one of the effective measures for copyright protection and the integrity of the relational database in the internet environment. In this paper, we propose a new watermarking scheme which can detect and localize the tampering of the relational database with categorical data. This watermarking scheme does not only change the values in the database, but also changes implicitly and securely the order of the tuples in the database.

Key words. Watermark, database, categorical data.

1. GIỚI THIỆU

Ngày nay, với sự phát triển mạnh mẽ của máy tính cùng sự bùng nổ của Internet đã giúp cho việc sử dụng và trao đổi thông tin càng ngày càng dễ dàng hơn. Tuy nhiên, nó cũng đem đến một nguy cơ đe dọa sự toàn vẹn của các thông tin khi được lưu thông trên Internet. Vì vậy, chủ sở hữu của những dữ liệu này luôn mong muốn dữ liệu của mình được phổ biến và phân phối một cách đúng đắn, không bị vi phạm bản quyền cũng như không bị giả mạo và xuyên tạc [1]. Đây là một yêu cầu rất cấp bách đòi hỏi phải có những công nghệ bảo vệ thật hữu hiệu. Trong đó, thủy vân đang nổi lên như một công nghệ mới có thể sử dụng để bảo vệ bản quyền cũng như bảo đảm sự toàn vẹn cho các cơ sở dữ liệu quan hệ [2, 4, 5, 7, 9, 15].

Gần đây, đã có một số lược đồ thủy vân bền vững được công bố nhằm bảo vệ bản quyền cho cơ sở dữ liệu quan hệ [1, 3, 5, 6, 8, 10, 14]. Tuy nhiên, việc bảo vệ bản quyền cho các cơ sở dữ liệu là tương đối tốn kém và không phải lúc nào các dữ liệu cũng cần phải chứng thực bản quyền mà có thể chỉ cần phải xác minh dữ liệu vẫn còn toàn vẹn là đủ. Tuy vậy, cho đến nay vẫn chưa có nhiều công trình nghiên cứu về khía cạnh này.

Để phát hiện giả mạo, các kỹ thuật thủy vân hiện có đều phải tiến hành những thay đổi nhỏ đối với giá trị của các thuộc tính trong một số bộ của cơ sở dữ liệu [1, 9, 10]. Tuy nhiên, có những dữ liệu không chấp nhận thay đổi cho dù rất nhỏ vì có thể sẽ làm mất ý nghĩa cũng như giá trị sử dụng của chúng. Để khắc phục điểm này, Y.Li, H Guo và S Jajodia [11] đã đề xuất một lược đồ thủy vân mới có thể khoanh vùng các giả mạo hoặc xuyên tạc đối với cơ sở dữ liệu quan hệ mà không làm thay đổi bất kỳ giá trị dữ liệu nào. Điểm mấu chốt của kỹ thuật này là việc đổi thứ tự của các bộ trong cơ sở dữ liệu. Tuy nhiên, lược đồ này có tính an toàn chưa cao do việc đổi thứ tự của các bộ chỉ được thực hiện trên các cặp đã định trước và bài báo chưa chứng minh được tính đúng đắn của thuật toán. Bài báo đề xuất một kỹ thuật đổi thứ tự mới bằng cách chọn các cặp bộ dựa vào một khóa bí mật K và số các bộ trong mỗi nhóm, nhằm tăng cường tính bảo mật thông qua các tham số bí mật chỉ người chủ dữ liệu biết và chứng minh tính đúng đắn của thuật toán cũng như thử nghiệm thuật toán.

Trong Mục 2 sẽ đưa ra một số thuật ngữ về thủy vân cơ sở dữ liệu quan hệ, khóa thủy vân, hàm hash và dữ liệu phân loại. Mục 3 trình bày thuật toán nhúng thủy vân và thuật toán phát hiện thủy vân đã nhúng trong quan hệ được cải tiến để nâng cao tính bảo mật từ thuật toán trong [11]. Mục 4 sẽ chứng minh tính đúng đắn của thuật toán. Mục 5 là cân đối giữa số bộ trong quan hệ và số nhóm. Cuối cùng là phần kết luận.

2. MỘT SỐ ĐỊNH NGHĨA

2.1. Thủy vân

Trong các nghiên cứu về các giải pháp bảo vệ bản quyền và sự toàn vẹn của các cơ sở dữ liệu quan hệ trong môi trường Internet, khái niệm thủy vân cơ sở dữ liệu luôn được nhắc đến [1-4, 7, 10, 11, 13-15], nhưng chưa được định nghĩa một cách thống nhất. Sau khi tổng hợp lại, chúng tôi đưa ra một định nghĩa chung nhất cho khái niệm này như sau.

Định nghĩa 1. Thủy vân cơ sở dữ liệu quan hệ là một kỹ thuật nhúng một số thông tin nào đó (được gọi là thông tin thủy vân W) vào cơ sở dữ liệu quan hệ nhằm mục đích bảo vệ bản quyền hoặc sự toàn vẹn cho cơ sở dữ liệu này. Thủy vân có thể ở dạng ẩn hoặc hiện và có thể là bền vững hoặc dễ vỡ.

2.2. Khóa thủy vân

Để chủ sở hữu của cơ sở dữ liệu có thể giữ bí mật cho thông tin thủy vân W và là người duy nhất có thể tìm lại được thông tin này thì cần phải trộn W với một thông tin đặc biệt được gọi là khóa do chính chủ cơ sở dữ liệu tự chọn. Thông tin thứ hai này được gọi là khóa thủy vân và được định nghĩa như sau.

Định nghĩa 2. Khóa thủy vân là một chuỗi các bit bí mật do chủ sở hữu cơ sở dữ liệu tự chọn (được gọi là K). Khóa K sẽ được trộn với thủy vân W để nhúng vào cơ sở dữ liệu. Sau đó, K sẽ đóng vai trò là khóa trong qui trình tìm lại thủy vân.

Một trong những cách giấu khóa hữu hiệu nhất là sử dụng hàm hash vì kỹ thuật này đảm bảo được yêu cầu bảo mật cũng như chi phí tính toán.

Định nghĩa 3. Hàm hash [12]

- Hàm hash là một thuật toán nhận vào một xâu ký tự có độ dài tùy ý và cho ra một xâu có độ dài qui định.

- Hàm hash mật mã học là một hàm hash với một số tính chất bảo mật nhất định để phù hợp với việc sử dụng trong nhiều ứng dụng bảo mật thông tin đa dạng (chứng thực và kiểm tra tính nguyên vẹn của thông điệp).

2.3. Dữ liệu phân loại

Giả sử ta có một bảng ghi lại danh sách các nhân viên trong một công ty, trong đó có một cột ghi số hiệu phòng làm việc của từng người. Sử dụng cột này, ta có thể chia các nhân viên theo số hiệu phòng. Vì thế thuộc tính này được gọi là thuộc tính phân loại. Giá trị của các thuộc tính phân loại không thể bị thay đổi một cách tùy tiện vì như vậy sẽ ảnh hưởng đến việc phân loại các đối tượng trong bảng dữ liệu. Dữ liệu phân loại trong cơ sở dữ liệu quan hệ có thể được định nghĩa như sau.

Định nghĩa 4. Dữ liệu phân loại trong cơ sở dữ liệu quan hệ là dữ liệu có thể được dùng để phân chia các bộ thành một số loại khác nhau. Một thay đổi nhỏ về giá trị của dữ liệu có thể làm ảnh hưởng đến giá trị sử dụng của toàn bộ dữ liệu trong cơ sở dữ liệu quan hệ.

3. LƯỢC ĐỒ NHÚNG VÀ PHÁT HIỆN THỦY VÂN

Trong Mục 2 đã đưa ra các định nghĩa cần thiết. Lược đồ thủy vân cơ sở dữ liệu quan hệ bao gồm 2 phần: nhúng thủy vân và phát hiện thủy vân [8]. Khi nhúng thủy vân, một khóa bí mật K do chủ sở hữu cơ sở dữ liệu tự chọn sẽ được sử dụng để nhúng thủy vân W vào cơ sở dữ liệu gốc. Sau khi nhúng thủy vân, các cơ sở dữ liệu sẽ được đưa vào lưu thông trong môi trường mạng. Để xác minh quyền sở hữu của một cơ sở dữ liệu đáng ngờ, quá trình xác minh được thực hiện mà cơ sở dữ liệu bị nghi ngờ được thực hiện như là đầu vào và bằng cách sử dụng khóa bí mật K (được sử dụng trong giai đoạn nhúng) thủy vân nhúng (nếu có) được lấy ra và so sánh với các thông tin thủy vân ban đầu.

Một vấn đề quan trọng trong một lược đồ thủy vân là sự đồng bộ hóa, nghĩa là phải đảm bảo thủy vân được trích ra theo đúng thứ tự đã được nhúng vào [8, 11]. Nếu mất sự đồng bộ này thì ngay cả khi không có bất kỳ một sửa đổi nào, thủy vân đã nhúng cũng có thể không được xác minh đúng đắn. Trong một lược đồ thủy vân dễ vỡ đối với các dữ liệu đa phương tiện, sự đồng bộ hóa không phải là vấn đề vì vị trí tương đối của các dữ liệu đa phương tiện là cố định [3]. Ngược lại, các bộ trong một quan hệ là độc lập với nhau và có thể được đặt theo một thứ tự tùy ý. Đây là một đặc điểm mà chúng ta có thể sử dụng để đưa ra một kỹ thuật mới cho lược đồ thủy vân đối với dữ liệu phân loại thông qua việc vận dụng thứ tự của các bộ. Ưu điểm của cách tiếp cận này là không thay đổi giá trị thuộc tính, một giải pháp lý tưởng khi thủy vân các dữ liệu phân loại, mà chỉ làm thay đổi thứ tự của các bộ trong quan hệ.

Trong lược đồ thủy vân đề xuất, tư tưởng chính của các thuật toán dựa vào [11]. Giống như lược đồ thủy vân [11], quá trình nhúng thủy vân và phát hiện thủy vân được thực hiện

trên từng nhóm. Mỗi nhóm này có thể được xem như là một nhóm ảo mà không thay đổi vị trí vật lý của các bộ. Sau khi nhóm, tất cả các bộ trong mỗi nhóm được sắp xếp theo khóa chính của nó. Giống như khi chia nhóm, việc sắp xếp này không làm thay đổi vị trí vật lý của các bộ. Mỗi nhóm sau đó sẽ được xử lý một cách độc lập.

Trong thuật toán nhúng thủy vân của [11], việc chọn các cặp bộ để nhúng thủy vân được thực hiện tuần tự cho nên rất dễ bị phát hiện. Khác với cách chọn này của [11], với mỗi cặp, chỉ lấy bộ thứ nhất theo thứ tự, còn bộ thứ hai của cặp bộ này được chọn dựa vào khóa nhúng thủy vân K và số bộ trong nhóm G_k. Quá trình chọn cặp để đổi chỗ này sẽ giúp tăng độ an toàn cho thủy vân được nhúng trong quan hệ. Đây là điểm mạnh của lược đồ thay cho lược đồ trong [11] việc đổi chỗ được tiến hành cho các bộ một cách cố định sẽ tạo cơ hội tấn công cao hơn từ những kẻ muốn sửa đổi quan hệ. Từ đó, việc đổi chỗ hai bộ không phải cố định thể hiện sự vượt trội hơn so với việc chỉ đổi chỗ cố định. Thêm vào nữa, việc đổi chỗ các cặp bộ dựa vào khóa bí mật K và tham số g, mà K và g chỉ chủ sở hữu dữ liệu biết sẽ tăng tính bảo mật cho quan hệ được nhúng. Từ đó, thực sự kẻ tấn công khó có thể tìm ra vị trí các cặp bộ được đổi chỗ cho nhau.

Giả sử có một quan hệ với một khóa chính là P, γ thuộc tính và có dữ liệu phân loại được ký hiệu là $R(P, A_1, A_2, ..., A_\gamma)$ và K là khóa thủy vân. Bảng 1 là các ký hiệu và tham số sẽ được sử dụng trong lược đồ thủy vân đề xuất.

Bảng 1. Các ký hiệu và các tham số

Ký hiệu	Ý nghĩa của ký hiệu
γ	Số thuộc tính của quan hệ
ω	Số bộ của quan hệ
g	Số nhóm của quan hệ
h_i	Giá trị hash của bộ thứ i trong quan hệ hoặc trong 1 nhóm
H_i	Giá trị hash của khóa thủy vân và khóa chính của bộ thứ i trong quan hệ hoặc trong 1 nhóm
H	Giá trị hash của nhóm
G_k	Nhóm thứ k
q_k	Số bộ trong nhóm thứ k
r_i	Bộ thứ i trong quan hệ hoặc trong một nhóm
$r_i.A_j$	Giá trị thuộc tính thứ j của bộ thứ i trong quan hệ hoặc trong một nhóm
K	Khóa nhúng thủy vân
W	Chuỗi thủy vân nhúng trong một nhóm
$W[i]$	Bít thủy vân thứ i trong chuỗi thủy vân W
W^*	Thủy vân được trích ra từ một nhóm
$W^*[i]$	Bít thủy vân thứ i trong chuỗi thủy vân W^*
V	Kết quả xác minh thủy vân

3.1. Nhúng thủy vân

Quá trình nhúng thủy vân vào quan hệ được thực hiện bằng Thuật toán 1 dưới đây.

Thuật toán 1. Nhúng thủy vân

1. **for** $k = 1$ **to** g **do** // khởi tạo các chỉ số và các nhóm
2. $q_k = 0$
3. $G_k = \emptyset$
4. **end for**
5. **for** $i = 1$ **to** ω **do** // chia các bộ thành g nhóm
6. $h_i = HASH(K, r_i.A_1, r_i.A_2, ..., r_i.A_\gamma)$ // giá trị hash của bộ thứ i
7. $H_i = HASH(K, r_i.p)$
8. $k = H_i \mod g$ // xác định nhóm
9. $r_i \rightarrow G_k$ // đưa bộ r_i vào nhóm G_k
10. q_k++
11. **end for**
12. **for** $k = 1$ **to** g **do**
13. Lưu khóa chính của các bộ trong G_k vào một cột trung gian T và sắp thứ tự
14. $H = HASH(K, h_{k_1}, h_{k_2}, ..., h_{k_{q_k}})$//$h_{k_i}$: giá trị hash của bộ nhóm k sau sắp thứ tự
15. $W = ExtractBits(H, \lfloor q_k/2 \rfloor)$ // $\lfloor x \rfloor$: phần nguyên của x
16. **for** $i = 1$ **to** $\lfloor q_k/2 \rfloor$ **do**
17. $j = HASH(K) \mod (q_k - 2i + 1) + 2$
18. **If** ($h_{k_1} \leq h_{k_j}$ **and** $W[i] == 1$) **or** ($h_{k_1} > h_{k_j}$ **and** $W[i] == 0$) **then**
19. đổi chỗ bộ r_{k_1} và r_{k_j} của quan hệ
20. **end if**
21. Loại bỏ khóa chính của bộ r_{k_1} và r_{k_j} ra khỏi cột T và sắp lại thứ tự
22. **end for**
23. ExtractBits(H, ℓ) {
24. **if** $length(H) > \ell$ **then**
25. W được gán bằng ℓ bit đầu tiên của H
26. **else**
27. $m = \ell - length(H)$
28. W được gán bằng H ghép với ExtractBits (H, m)
29. **end if**
30. **return** W **}**
31. **end for**

Giải thích Thuật toán 1: Đầu tiên, các bộ của quan hệ được chia thành g nhóm theo khóa chính và khóa nhúng thủy vân K. Việc chia nhóm này chỉ là một hoạt động ảo, không làm

thay đổi vị trí vật lý của các bộ trong quan hệ. Đồng thời, tính các giá trị hash h_i của các thuộc tính (trừ khóa chính) của bộ r_i ghép với khóa thủy vân K, trong đó mỗi giá trị thuộc tính là một chuỗi bit.

Tiếp theo là việc nhúng thủy vân vào từng nhóm. Thủy vân là một chuỗi W gồm $\lfloor q_k/2 \rfloor$ bit được trích ra bằng hàm $ExtractBits(H, \lfloor q_k/2 \rfloor)$ từ chuỗi bít H, trong đó H là giá trị hash của nhóm được tính từ các giá trị h_i ghép với khóa K. Việc nhúng thủy vân thực chất là việc đổi chỗ hai bộ dựa trên giá trị của chuỗi bit thủy vân W và giá trị hash của các bộ này. Các cặp bộ được chọn phụ thuộc vào giá trị hash khóa thủy vân K và số bộ trong nhóm. Sau đó cặp được chọn có thể bị đổi chỗ hay không tùy thuộc vào bít thủy vân W của cặp.

3.2. Phát hiện thủy vân

Để kiểm tra tính toàn vẹn của quan hệ cũng như phát hiện xem có các giả mạo nào trong quan hệ đã nhúng thủy vân hay không thì cần phải phát triển một thuật toán để tìm lại thủy vân. Thuật toán 2 sau đây được sử dụng để phát hiện thủy vân đã được nhúng trong quan hệ bằng Thuật toán 1. Thuật toán này cần phải biết khóa thủy vân Kvà số nhóm g.

Thuật toán 2. Phát hiện thủy vân

1. **for** $k = 1$ **to** g **do** // khởi tạo các chỉ số và các nhóm
2. $q_k = 0$
3. $G_k = \emptyset$
4. **end for**
5. **for** $i = 1$ **to** ω **do** // chia các bộ thành g nhóm
6. $h_i = HASH(K, r_i.A_1, r_i.A_2, ..., r_i.A_\gamma)$
7. $H_i = HASH(K, r_i.p)$
8. $k = H_i \bmod g$ // xác định nhóm
9. $r_i \rightarrow G_k$ // đưa bộ r_i vào nhóm G_k
10. q_k++
11. **end for**
12. **for** $k = 1$ **to** g **do**
13. Lưu khóa chính của các bộ trong G_k vào một cột trung gian T và sắp thứ tự
14. $H = HASH(K, h_{k_1}, h_{k_2}, ..., h_{k_{q_k}})$// h_{k_i}: giá trị hash của bộ nhóm k sau sắp thứ tự
15. $W = ExtractBits(H, \lfloor q_k/2 \rfloor)$ // $\lfloor x \rfloor$: phần nguyên của x
16. **for** $i = 1$ **to** $\lfloor q_k/2 \rfloor$ **do**
17. $j = HASH(K) \bmod (q_k - 2i + 1) + 2$
18. **If** $h_{k_1} < h_{k_j}$ **then**
19. $W^*[i] = 0$
20. **else**

21. $W^*[i] = 1$
22. **end if**
23. Loại bỏ khóa chính của bộ r_{k_1} và r_{k_j} khỏi cột T và sắp lại thứ tự
24. **end for**
25. **if** $W^* == W$ **then**
26. V = true
27. **else**
28. V = false
29. **end if**
30. **end for**

Giải thích Thuật toán 2: Phát hiện thủy vân trong từng nhóm được thực hiện dựa trên việc so sánh 2 chuỗi thủy vân. Chuỗi thứ nhất được tính tương tự như Thuật toán 1. Chuỗi thứ hai được rút ra từ quan hệ cần xác minh bằng cách so sánh giá trị hash của các cặp bộ trong nhóm.

Nếu như kết quả so sánh hai chuỗi là trùng nhau thì ta kết luận không có giả mạo, ngược lại thì có giả mạo trong nhóm đang xét. Nếu tất cả các nhóm không có giả mạo thì kết luận quan hệ đem xác minh là toàn vẹn. Ngược lại, quan hệ là không toàn vẹn và xác minh các nhóm là không toàn vẹn.

Từ thuật toán nhúng thủy vân và phát hiện thủy vân, có thể dễ dàng nhận thấy độ phức tạp tính toán của cả hai thuật toán có bậc bằng số các bộ trong quan hệ, tức là bậc $O(\omega)$.

4. TÍNH ĐÚNG ĐẮN

Trong phần này sẽ chứng minh tính đúng đắn của các thuật toán đề xuất. Để chứng minh tính đúng đắn của định lý, ta có mệnh đề sau.

Mệnh đề 1. *Nếu trong thuật toán nhúng, một bộ thuộc vào nhóm G_k nào đó và giá trị khóa chính của nó không bị thay đổi thì bộ này vẫn thuộc nhóm G_k trong thuật toán phát hiện thủy vân.*

Chứng minh. Giả sử có một quan hệ có ω bộ với một khóa chính là P, γ thuộc tính và có dữ liệu phân loại, ký hiệu là $R(P, A_1, A_2, ..., A_\gamma)$. Xét bộ $r_i(p, a_1, a_2, ..., a_\gamma)$ trong quan hệ R. Theo giả thiết, khóa chính p không bị thay đổi khi quan hệ bị tấn công và chủ sở hữu quan hệ hoặc người có thẩm quyền biết về khóa bí mật K và số nhóm phân chia g.

Trong thuật toán nhúng thủy vân, ta có:

+ Theo tính chất của hàm HASH;

+ $H_i = HASH(K, r_i.p)$ chỉ phụ thuộc vào giá trị các tham số K và $r_i.p$;

+ Chỉ số nhóm $k = H_i \mod g$;

$\Longrightarrow r_i \in G_k$.

Trong thuật toán phát hiện thủy vân, theo giả thiết ta có $r_i.p$ không thay đổi và tham số

K, g đã biết trước và không thay đổi. Khi đó:

+ Theo tính chất của hàm HASH;

+ $H_i = HASH(K, r_i.p)$ chỉ phụ thuộc vào giá trị các tham số K và $r_i.p$;

+ Mặt khác, việc xác định chỉ số nhóm của r_i theo công thức $k = H_i \mod g$;

+ Tham số $g, K, r_i.p$ không thay đổi;

$\Longrightarrow r_i \in G_k$.

Do đó, $r_i \in G_k$ trong cả hai thuật toán nhúng và phát hiện thủy vân. Mà r_i là một bộ được chọn ngẫu nhiên trong quan hệ R. Vì vậy, ta có điều phải chứng minh. ■

Định lý 1. *Nếu một quan hệ có dữ liệu phân loại không bị giả mạo hoặc xuyên tạc thì thủy vân đã nhúng vào quan hệ bằng Thuật toán 1 và thủy vân được trích ra từ quan hệ bằng Thuật toán 2 là như nhau.*

Chứng minh. Giả sử có một quan hệ có ω bộ với một khóa chính là P, γ thuộc tính và có dữ liệu phân loại, ký hiệu là $R(P, A_1, A_2, ..., A_\gamma)$. Theo giả thiết, quan hệ không bị thay đổi. Vì vậy, theo Mệnh đề 1 ω bộ được phân vào g nhóm $G_i (i = 1, 2, ..., g)$ là giống nhau trong thuật toán nhúng và phát hiện thủy vân.

Giả sử xét nhóm G_k, có q_k bộ. $W(W[1], W[2], ..., W[\lfloor q_k/2 \rfloor])$ là thủy vân được sinh ra trong quá trình nhúng thủy vân, trong đó

$$W[i] = \{0, 1\},\ i = 1, 2, ..., \lfloor q_k/2 \rfloor;\ W'(W'[1], W'[2], ..., W'[\lfloor q_k/2 \rfloor])$$

là thủy vân được sinh ra trong quá trình phát hiện thủy vân, trong đó $W'[i] = \{0, 1\}$. $W^*(W^*[1], W^*[2], ..., W^*[\lfloor q_k/2 \rfloor])$ là thủy vân được trích ra từ quan hệ trong thuật toán phát hiện thủy vân, trong đó $W^*[i] = \{0, 1\}$. Vì lược đồ thủy vân sử dụng duy nhất một cách sắp thứ tự cho cả thuật toán nhúng và phát hiện thủy vân cho nên các bộ r_i $(i = 1, 2, ..., q_k)$ được sắp thứ tự như nhau trong phần nhúng và phát hiện thủy vân.

Trong phần nhúng, theo tính chất của hàm HASH, và cách tính:

+ $h_i = HASH(K, r_i.A_1, r_i.A_2, ..., r_i.A_\gamma)$;

+ $H = HASH(K, h'_1, h'_2, ..., h'_{q_k})$;

trong đó, h'_j là giá trị h_i sau khi sắp thứ tự. Sử dụng hàm $ExtractBits$ với tham số đầu vào là giá trị của H ta sẽ sinh ra được chuỗi thủy vân nhúng W:

$$W = ExtractBits(H, \lfloor q_k/2 \rfloor).$$

Trong phần phát hiện, tương tự như trên ta có:

+ $h'_i = HASH(K, r_i.A_1, r_i.A_2, ..., r_i.A_\gamma)$;

+ $H' = HASH(K, h'_1, h'_2, ..., h'_{q_k})$;

+ $W' = ExtractBits(H', \lfloor q_k/2 \rfloor)$;

+ Vì quan hệ không bị giả mạo hoặc xuyên tạc và cũng được sắp thứ tự như nhau cho nên các $r_i.A_j$ là không thay đổi trong phần phát hiện.

$$\Longrightarrow h'_i = h_i\ \forall i = 1, 2, ..., q_k \Rightarrow H' = H \Leftrightarrow W' = W. \quad (1)$$

Xét thuật toán nhúng thủy vân:

If ($h_{k_1} \leq h_{k_j}$ **and** $W[i] == 1$) **or** ($h_{k_1} > h_{k_j}$ **and** $W[i] == 0$) **then**

đổi chỗ bộ r_{k_1} và r_{k_j} của quan hệ

Như vậy, ta xét các điều kiện để đổi chỗ bộ r_{k_1} và r_{k_j}:

+ $h_{k_1} < h_{k_j}$ và $W[i] = 0$ thì không đổi chỗ.

+ $h_{k_1} < h_{k_j}$ và $W[i] = 1$ thì đổi chỗ. Sau đổi chỗ thì $h_{k_1} > h_{k_j}$.

+ $h_{k_1} \geq h_{k_j}$ và $W[i] = 0$ thì đổi chỗ. Sau đổi chỗ thì $h_{k_1} \leq h_{k_j}$.

+ $h_{k_1} \geq h_{k_j}$ và $W[i] = 1$ thì không đổi chỗ.

$\Leftrightarrow h_{k_1} \geq h_{k_j}$ thì $W[i] = 1$ và $h_{k_1} < h_{k_j}$ thì $W[i] = 0$. (2)

Các bit $W[i]$ của thủy vân W lần lượt được nhúng vào trong nhóm G_k bằng việc đổi chỗ các bộ trong quan hệ ban đầu như đã xét trong các trường hợp ở trên.

Tương tự, ta xét quá trình trích thủy vân W^* trong thuật toán phát hiện:

If $h_{k_1} < h_{k_j}$ **then**

$W^*[i] = 0$ (3)

else

$W^*[i] = 1$

Từ (3) $\Longrightarrow h_{k_1} < h_{k_j}$ thì $W^*[i] = 0$ và khi $h_{k_1} \geq h_{k_j}$ thì $W^*[i] = 1$. (4)

Từ (2), (4) và do quan hệ không bị thay đổi cho nên $W^*[i] = W[i]$. Vì $W^*[i]$, $W[i]$ là bất kỳ nên $W^* = W$. (5)

Lược đồ thủy vân thực hiện trên nhóm $G_k (k = 1, 2, ..., g)$ và các nhóm độc lập với nhau. (6)

Từ (1), (5) và (6), ta có điều phải chứng minh. ■

5. CÂN ĐỐI GIỮA SỐ BỘ TRONG QUAN HỆ VÀ SỐ NHÓM

Với lược đồ đề xuất ở trên, nếu có một nhóm bị phát hiện là giả mạo thì người ta có thể loại nhóm đó đi và các nhóm không bị giả mạo sẽ tiếp tục được sử dụng. Việc xác định một nhóm nào đó có bị giả mạo hay không có thể được thực hiện trong Thuật toán 2.

Trước khi thủy vân một quan hệ, cần phải lựa chọn tham số g (số nhóm thủy vân) sao cho phù hợp với số bộ trong quan hệ. Số các bộ của quan hệ và số nhóm phải được chọn như thế nào đó để có thể thỏa mãn đồng thời hai tính chất. Đó là tăng cường tính bền vững của thủy vân và tối đa số các bộ có thể tiếp tục được sử dụng. Có thể thấy ngay là không thể nào thỏa mãn được đồng thời hai tính chất này. Điều này được khẳng định bằng các Mệnh đề 2 và Mệnh đề 3. Vì vậy, sẽ phải có một sự thỏa hiệp để cân đối giữa hai tính chất trên. Số lượng nhóm nên được chọn vừa đủ tương ứng với số lượng các bộ của quan hệ để vừa có các chuỗi thủy vân bền vững trong mỗi nhóm và vừa có thể tiếp tục sử dụng được nhiều bộ nhất. Nhưng nếu cần phải tăng cường tính bền vững của thủy vân thì nên chọn g nhỏ để số bộ trong mỗi nhóm sẽ tăng lên cùng chiều với độ bền vững. Nếu ngược lại, tức là nhu cầu tiếp tục sử dụng những bộ không bị xâm phạm là cấp bách thì cần phải chọn g lớn để số lượng

các bộ phải loại bỏ sẽ ít hơn. Trong các thử nghiệm, với các quan hệ có khoảng 10.000 bộ cho thấy g nằm trong khoảng từ 6 đến 10 là những lựa chọn tốt cho cả hai tính chất đã đề cập trên đây.

Mệnh đề 2. *Cho một quan hệ có dữ liệu phân loại được thủy vân bằng Thuật toán 1 và 2. Nếu quan hệ R có kích thước không đổi và số nhóm g tăng thì:*

1. Số lượng bộ có thể tiếp tục sử dụng dữ liệu tăng.

2. Độ bền vững của thủy vân giảm.

Chứng minh. Giả sử có một quan hệ có ω bộ với một khóa chính là P, γ thuộc tính và có dữ liệu phân loại, ký hiệu là $R(P, A_1, A_2, ..., A_\gamma)$. Theo giả thiết số bộ ω là cố định.

1. Chứng minh số lượng bộ có thể tiếp tục sử dụng dữ liệu tăng.

Theo giả thiết ta có:

+ ω bộ được phân vào g nhóm $G_k(k = 1, 2, ..., g)$ bằng thuật toán nhúng và phát hiện thủy vân.

+ Số bộ ω là cố định.

Theo Thuật toán 1 và 2, số bộ trung bình trong mỗi nhóm G_k là $q_k(q_k = \omega/g)$.

Khi g tăng lên thì số bộ trung bình trong mỗi nhóm G_k là β_k. Ta dễ dàng nhận thấy $\beta_k < q_k(\omega/g' < \omega/g$, với g' là số nhóm sau khi tăng).

Giả sử quan hệ có sửa đổi xảy ra và không mất tổng quát, dựa vào Thuật toán 2, phát hiện nhóm G_k bị sửa đổi. Hay số lượng bộ có thể tiếp tục sử dụng dữ liệu của quan hệ R sẽ loại đi số bộ trong nhóm G_k. Dễ dàng nhận thấy số lượng bộ dữ liệu trung bình bị loại đi trong G_k sau khi g tăng nhỏ hơn trước khi tăng g (giảm đi $q_k - \beta_k$ bộ). Khi đó số lượng bộ có thể tiếp tục sử dụng dữ liệu tăng lên $q_k - \beta_k$ bộ.

2. Chứng minh độ bền vững của thủy vân giảm.

Ta có thể thấy, độ bền vững của thủy vân trong Thuật toán 1 và 2 dựa vào độ dài chuỗi thủy vân W. Như vậy sẽ chứng minh độ bền vững dựa vào độ dài của W.

Theo giả thiết ta có: g tăng $\Longrightarrow$ số lượng bộ dữ liệu trung bình trong nhóm G_k giảm, suy ra chuỗi thủy vân W và chuỗi thủy vân trích ra W^* của G_k có độ dài sẽ giảm đi.

Suy ra điều phải chứng minh. ■

Mệnh đề 3. *Cho quan hệ R có dữ liệu phân loại được thủy vân bằng Thuật toán 1 và 2 với số nhóm không đổi, nếu như kích thước quan hệ tăng thì:*

1. Số lượng bộ có thể tiếp tục sử dụng dữ liệu giảm.

2. Độ bền vững của thủy vân tăng.

Chứng minh. Giả sử có một quan hệ có ω bộ với một khóa chính là P, γ thuộc tính và có dữ liệu phân loại, ký hiệu là $R(P, A_1, A_2, ..., A_\gamma)$. Theo giả thiết g cố định.

1. Chứng minh: Số lượng bộ có thể tiếp tục sử dụng dữ liệu giảm.

Theo giả thiết ta có:

+ ω bộ được phân vào g nhóm G_k (k = 1, 2, ..., g) bằng thuật toán nhúng và phát hiện

thủy vân.

+ Số nhóm g là cố định nên số phần tử trong nhóm G_k là q_k.

Khi ω tăng lên thì số lượng phần tử trong G_k là $q_k + \beta_k$ (với β_k là số bộ tăng thêm của G_k khi ω tăng).

Giả sử quan hệ có sửa đổi xảy ra và không mất tổng quát, dựa vào Thuật toán 2, phát hiện nhóm G_k bị sửa đổi. Khi đó, số lượng bộ có thể tiếp tục sử dụng dữ liệu của quan hệ R sẽ loại đi số bộ trong nhóm G_k. Hay số lượng bộ bị loại đi trong G_k sau khi ω tăng lớn hơn trước khi tăng ω là β_k bộ. Suy ra, số lượng bộ có thể tiếp tục sử dụng dữ liệu giảm đi β_k bộ.

2. Chứng minh: Độ bền vững của thủy vân tăng.

Ta có thể thấy, độ bền vững của thủy vân trong Thuật toán 1 và 2 dựa vào độ dài chuỗi thủy vân W. Như vậy sẽ chứng minh độ bền vững dựa vào độ dài của W.

Theo giả thiết ta có: ω tăng, suy ra số lượng bộ dữ liệu trong nhóm G_k tăng $\Longrightarrow$ chuỗi thủy vân W và chuỗi thủy vân trích ra W^* của G_k có độ dài sẽ tăng lên.

Suy ra điều phải chứng minh. ■

6. KẾT LUẬN

Lược đồ thủy vân đề xuất làm việc trên các nhóm trong quan hệ của cơ sở dữ liệu quan hệ. Khóa thủy vân và tham số g (số lượng nhóm) được nhúng vào cơ sở dữ liệu quan hệ là một dữ liệu mật. Trong lược đồ này cần tạo ra sự cân bằng giữa độ an toàn và chi phí tính toán.

Kích thước của cơ sở dữ liệu càng lớn thì độ bền vững của lược đồ càng tốt nhưng ngược lại số lượng bộ có thể tiếp tục sử dụng dữ liệu giảm và chi phí tính toán lớn. Trong khi đó, số nhóm thủy vân càng lớn thì số lượng bộ có thể tiếp tục sử dụng dữ liệu càng tốt nhưng độ bền vững của lược đồ thủy vân giảm. Vì vậy trong lược đồ này việc chọn g được xem xét để cân đối với kích thước của quan hệ.

Lược đồ đề xuất sử dụng các nhóm độc lập trong quan hệ, có những điểm mạnh sau:

- Nhúng thủy vân không làm thay đổi giá trị của các bộ.

- Phát hiện và khoanh vùng giả mạo trên từng nhóm độc lập. Do vậy, các nhóm còn lại trong cơ sở dữ liệu vẫn còn có thể được sử dụng nếu cần thiết.

TÀI LIỆU THAM KHẢO

[1] A. Hamadou, X. Sun, S. A. Shah, L. Gao, A weight-based semi-fragile watermarking scheme for integrity verification of relational data, *International Journal of Digital Content Technology and its Applications* **5** (8) (2011).

[2] Agrawal, Kiernan, Haas, Watermarking relational data: framework, algorithms amd analysis, *The VLDB Journal* **12** (2) (2003) 157–169.

[3] Collberg and Thomborson, Watermarking, tamper- proofing, and obfuscation - tools for software protection, *IEEE Trans. Software Engrg* **28** (8) (2002) 735–746.

[4] P. Cousot, R. Cousot, An abstract interpretation -based framework for software watermarking, *31st ACM SIGPLANSIGACT Symposium On Principles Of Programming Languages, POPL'04*, Venice, Italy, January 14-16, 2004.

[5] H. M. Sardroudi, S. Ibrahim, O. Zanganeh, Robust database watermarking technique over numerical data, *JCIS* **1** 1 (2011) 30–40.

[6] J. Lafaye, An analysis of database watermarking security, *3rd International Symposium on Information Assurance and Security*, Manchester, United Kingdom, August 29-31, 2007 (IEEE Computer Society 2007).

[7] R. K. Bedi, P. Gujarathi, P. Gundecha, A unique approach for watermarking non-numeric relational database, *International Journal of Computer Applications* **36** (7) (2011) (0975–8887).

[8] Raju Halder, Shantanu Pal, Agostino Cortesi, Watermarking techniques for relational databases: survey, classification and comparison, *Journal of Universal Computer Science* **16** (21) (2010) 3164–3190.

[9] Sion, Proving ownership over categorical data, *Proceedings of the IEEE International Conference on Data Engineering IEEE ICDE 2004*, Boston, Massachusetts, March 2004 (584–596).

[10] Sukriti Bhattacharya, Agostino Cortesi, A distortion free watermark framework for relationnal databases, *Proc. 4th International Conference on Software and Data technology (ICSOFT) (2)*, Sofia, Bulgaria, 2009 (229–234).

[11] Yingjiu Li, Huiping Guo, Sushil Jajodia, Tamper detection and localization for categorical data using fragile watermarks, *DRM '04: Proceedings of the 4th ACM Workshop on Digital Rights Management*, New York, NY, USA, 2004 (73–82).

[12] Phan Đình Diệu, *Lý thuyết mật mã và an toàn thông tin*, NXB Đại học Quốc gia Hà Nội, 2002.

[13] Bùi Thế Hồng, Nguyễn Thị Thu Hằng, Lưu Thị Bích Hương, Thủy vân cơ sở dữ liệu quan hệ, *Tạp chí Khoa học và Công nghệ Đại học Thái Nguyên* **52** (4) (2009) 56–59.

[14] Bùi Thế Hồng, Lưu Thị Bích Hương, Thủy vân cơ sở dữ liệu quan hệ bằng kỹ thuật tối ưu, *Kỷ yếu hội thảo Quốc gia lần thứ XII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Đồng Nai, 05-06 tháng 8 năm 2009, (NXB Khoa học và Kỹ thuật, Hà Nội (2010) 443–457).

[15] Lưu Thị Bích Hương, Bùi Thế Hồng, Bảo vệ bản quyền công khai cho các cơ sở dữ liệu, *Kỷ yếu hội thảo Quốc gia lần thứ XIII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Hưng Yên, 19-20 tháng 8 năm 2010, (NXB Khoa học và Kỹ thuật, Hà Nội (2011) 41–50).

Ngày nhận bài 23 - 8 - 2012

Ngày lại sau sửa ngày 07 - 4 - 2013